关联数据中 owl:sameAs 网络分析*

贾君枝 李 晓
(山西大学经济与管理学院 太原 030006)

摘要: 【目的】调研 owl:sameAs 连接在真实数据网络中的配置和应用情况。【方法】从 BTC 2014 数据集中抽取部分数据, 应用统计学方法对样本数据构成的 sameAs 网络进行结构分析、域名分析和实例类型分析。【结果】数据分析结果表明, 真实数据网络中 sameAs 连接较稀疏, 等同实体网络中大多数实体只建立了单个连接。【局限】样本数据数量有限, 未能更全面地深入分析。【结论】该研究分析可以为关联数据中基于实例水平的数据集成、本体对齐、知识发现以及跨数据集查询等提供参考。

关键词: owl:sameAs 数据集互联 网络
分类号: G254
DOI: 10.11925/infotech.2096-3467.2017.0366

1 引 言

关联数据(Linked Data)建立在网络标准技术如HTTP、RDF 和 URIs 的基础上[1], 通过发布结构化数据, 建立数据相互连接而实现数据的增值。数据网络(Web of Data)是关联数据集的集合, 也称为关联开放数据(Linked Open Data)。2017 年 2 月 20 日, W3C 关联开放数据项目发布最新的关联开放数据云图(Linked Open Data Cloud, LOD Cloud), LOD 云图中包含的关联数据集已经由 2007 年 5 月的 12 个数据集增长到 1 139 个, 内容涵盖地理、政府、生命科学、语言、媒体、出版物、社交网络、用户生成内容 9 个领域[2]。在过去的几年中, 越来越多的社区将其数据以关联数据的形式发布到 LOD 中, 这种快速增长使得 LOD 云图成为知识发现和自动问答等应用的实验平台。数据发布者根据自身需要采用不同词表术语或自定义术语表示数据, 对于现实世界中的同一实体对象, 不同的数据发布者从自身角度出发从不同维度赋予其标识并进行描述, 从而增加了数据共享的难度。从不同数据集发现同一实体, 可以提高数据的互操作性。因此, 识别不同数据集的相同实体已成为数据的关联问题之一, 被人们关注并研究。拥有不同标识或 URIs 两个实体对象, 通过实例级关系 owl: sameAs 彼此连接。有研究表明, 数据集资源之间最重要的连接谓词之一是owl:sameAs[3]。找出不同数据集中基于 owl:sameAs 语义的实例也被定义为"实例对齐"[4]。

近几年, 本体对齐被认为是 LOD 中最重要的研究问题之一, 它是数据集成、跨数据集查询及知识获取的前提条件。在 LOD 环境中, 本体对齐主要包括三个部分: 概念(类)对齐, 属性对齐和实例对齐[4]。本体对齐中的很多研究基于实例之间的 owl:sameAs 连接展开。Parundekar 等[5]提出, 识别属于概念的等同(基于 owl:sameAs 连接)实例会导致这些概念之间的对齐。Correndo 等[6]在对齐概念中采用一种利用实例之间的 owl: sameAs 连接以及 Jaccard 系数测量实例重叠的统计学方法。Nikolov 等[7]利用 owl: sameAs 连接推断 LOD 中的本体概念之间的映射。Gunaratna 等[8]提出一种可以在 LOD 环境中使用的属性对齐的方法, 利用数据实例之间的现有实体共现链接(如使用 owl:sameAs 和 skos:closeMatch 形式的链接)匹配属性扩展。

为了建立更多的外部关联, 数据发布者通过一些自动和半自动的方法发现网络中的等同实体, 并建立

通讯作者: 贾君枝, ORCID: 0000-0003-1486-673X, E-mail: junzhij@163.com。
*本文系国家社会科学基金重点项目"基于关联数据的中文名称规范档语义描述及数据聚合研究"(项目编号: 15ATQ004)的研究成果之一。

owl:sameAs 连接。因此，伴随着数据网络的急速增长，跨数据集实例之间的 owl:sameAs 连接数量也在增长。虽然单个 owl:sameAs 谓词仅连接两个资源，但当数据网络中所有的 owl:sameAs 谓词及其连接的 RDF 资源汇聚在一起时，就形成一张巨大的有向图，称为 sameAs 网络。本文对真实数据网络中的 sameAs 网络作统计学分析，以期得到跨数据集之间实例的 owl:sameAs 配置和使用情况，为关联数据中基于实例水平的数据集成、本体对齐、跨数据集查询以及知识发现等研究提供参考。

2 owl:sameAs 特性

owl:sameAs 是万维网本体语言(OWL)的一个内建属性，用于将两个个体连接在一起。事实上要求每个人都使用相同的名字指称同一个个体是不现实的。当两个不同 URI 参引实际指的是同一个事物时，可以通过属性 owl:sameAs 将它们相连，表明被连接的两个个体有相同的"身份"[9]。比如，可以通过以下陈述表示两个 URI 参引实际指的是同一个人:

```
<rdf:Descriptionrdf:about="#William_Jefferson_Clinton">
<owl:sameAs rdf:resource="#BillClinton"/>
</rdf:Description>
```

假设拥有不同 URL 的两个个体是相同的实体，或者单个个体拥有多个名字，可以通过 owl:sameAs 属性声明它们的同一性关系。owl:sameAs 广泛应用于关联数据集中，通过可参引的 HTTP URL 提供了可以指向外部"等价"资源的可选方式，URL 自身可以唯一识别远程文档中的匹配资源。owl:sameAs 陈述经常用来定义本体之间的映射[9]。在关联数据社区中，由于 owl:sameAs 可以连接分布式数据集中的相同资源，因此它经常被用来支持关联数据聚合。

sameAs 陈述: 是指由 owl:sameAs 谓词连接两个 RDF 资源构成的三元组。其中，两个资源及谓词都由可参引的 HTTP URL 作为标识符。如下所示为一个 sameAs 陈述:

```
<http://data.linkedmdb.org/resource/film/13508>
<http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/The_Temptress>.
```

sameAs 网络: 网络从图论意义上理解是指由节点和连线构成的图，可以用带箭头的连线表示从一个节点到另一个节点存在的某种顺序关系[10]。把数据网络中所有 sameAs 陈述中的 RDF 资源表示成节点，用有方向的连线表示 owl:sameAs 关系，由此形成的网络称之为 sameAs 网络。

3 数据采集和分析

为获得真实数据网络中 owl:sameAs 的使用情况，本文选择的数据来源于 Billion Triple Challenge (BTC) 2014 数据集[11]。BTC 2014 数据集对网络数据的覆盖比率很高，其使用包括 VOID 描述和数据管理系统 CKAN 所有示例 URIs 在内的众多数据源作为种子集合，在网络中爬行了近 5 个月，截止到 2014 年 6 月，共采集 4 090 758 596 个 RDF 三元组，其中包含大量的 sameAs 陈述。本文从该数据集中抽取了 4 个数据包，共计 2 096 904 个三元组，进行处理和分析，使用的数据处理工具主要是 SQL Server。通常假设顶级域名相同的数据来自同一个数据集。为了获得真实数据网络中不同数据集之间的互联方式，对数据进一步处理，首先去掉无效和重复记录，然后提取主体和客体资源的顶级域名，从而得到主体和客体资源分别来自不同数据集的三元组共有 190 549 个。基于实例的数据集之间通过不同的谓词实现互联。对谓词进行统计，筛选出 URI 有效链接并且为多个数据集之间通用的谓词，如表 1 所示，可以看出 owl:sameAs 连接为数据集互联做出了巨大的贡献。

190 549 个三元组中有 45 846 个 sameAs 陈述。统计这些 sameAs 陈述中用于表示 owl:sameAs 属性的谓词形式及数量，如表 2 所示。可知，在数据网络中绝大多数 sameAs 陈述都使用了<http://www.w3.org/2002/07/owl#sameAs>这一规范的表达形式表示 sameAs 属性。

另外，由于 Wikipedia 有多个语言版本，基于 Wikipedia 的 DBpedia 数据集也具备多语言知识库特性，目前可支持多达 92 种语言。DBpedia 中的资源与它的各个语言版本下对应资源也建立了大量的 sameAs 连接，类似这样的陈述总共有 3 505 条。这部分数据对研究意义不大，因此从 45 846 个 sameAs 陈述中把上述 3 505 条移除，最终得到 42 341 条 sameAs 陈述，其主、客体资源来自不同的数据集。笔者将这 42 341 个 sameAs 陈述形成的集合称为样本数据集，由之形成的 sameAs 网络称为样本 sameAs 网络。

表 1　数据集间互联统计

谓词缩写	谓词 URI 及备注	数量
rdf:type	<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> 定义实例和类之间的联系	64 449
owl:sameAs	<http://www.w3.org/2002/07/owl#sameAs> 表示由不同 URI 标识的两个 RDF 资源指的是同一个对象	44 746
skos:exactMatch	<http://www.w3.org/2004/02/skos/core#exactMatch> 连接两个有足够的可信度并在信息检索应用程序较大范围可以交替使用的概念，是 skos:closeMatch 的子属性	13 102
rdfs:seeAlso	<http://www.w3.org/2000/01/rdf-schema#seeAlso> 将一个资源关联到另一个解释它的资源	5 570
skos:closeMatch	<http://www.w3.org/2004/02/skos/core#closeMatch> 连接两个足够相似以致在一些信息检索应用程序可以交替使用的概念	1 490
dcterms:type	<http://purl.org/dc/terms/type> 描述文件格式、物理媒介或资源的维度	1 170

表 2　owl:sameAs 谓词表达形式

sameAs 谓词	数量	占比
<http://www.w3.org/2002/07/owl#sameAs>	44 746	97.60%
<http://www.w3.org/2000/01/rdf-schema#sameAs>	631	1.38%
<owl:sameAs>	445	0.97%
<htpp://www.abes.fr/owlsameAs>	16	0.03%
<http://lexvo.org/ontology#nearlySameAs>	4	0.009%
<http://linkedgeodata.org/ontology/gadmSameAs>	4	0.009%

3.1　sameAs 网络特征分析

sameAs 网络具有一定的网络拓扑结构，通过对其结构的分析可以研究 owl:sameAs 在真实数据网络中的配置情况。

(1) 连接组件规模

样本数据集包含 80 521 个无重复的 URI 资源和 42 341 个唯一性 sameAs 陈述，把 RDF 资源表示成节点，用有向线表示 owl:sameAs 关系，因此形成的 sameAs 网络中共有 80 521 个节点和 42 341 条有向边。

在网络 G=(V,E)中(其中 V 称为节点集，E 称为边集)，如果有一部分图 G'=(V', E'), V'是 V 的子集，E'是 E 的子集，且 E'中的任意一条边 e'_{ij} 必定与 E 中的边 e_{ij} 对应，则称 G'是 G 全部的子图。若一个图 G 中的每对不同节点 v_i, v_j 之间都至少存在一条简单路径，则称该图 G 是连通的。连通子图指网络 G 中的一个子图，在这个子图中任意两个节点之间都至少存在一条简单路径[10]。如果把一个有向图的所有有向边用无向边替换掉后生成一个连通(无向)图，这个有向图就被称为是弱连接。一个无向图 G 的最大连通子图被称为一个连接组件。一个网络可能存在多个连接组件。图 1 为样本 sameAs 网络中的一个连接组件，其规模(用节点个数表示)为 20。样本 sameAs 网络的所有连接组件规模分布如图 2 所示，横坐标为连接组件规模(用节点个数表示)，纵坐标为对应规模的连接组件个数。该网络中共有 38 558 个连接组件，每个连接组件平均覆盖了 2.1 个 URI 资源。97%的连接组件规模为 2，即只在两个 RDF 资源间建立了 owl:sameAs 连接。大型连接组件较少，有个别连接组件包含上百个 RDF 资源。大规模连接组件呈现星型拓扑结构，由较少节点作为中心节点，将来自其他数据集的“等同”资源聚集在一起。因此，样本 sameAs 网络的典型尺寸是一个比较小的常数，并且规模增长也是比较缓慢的。

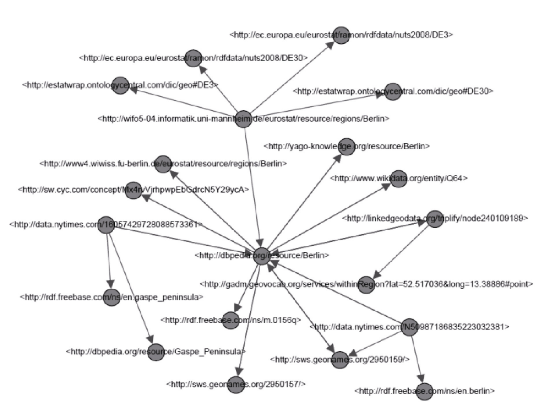


图 1　样本 sameAs 网络中的一个连接组件

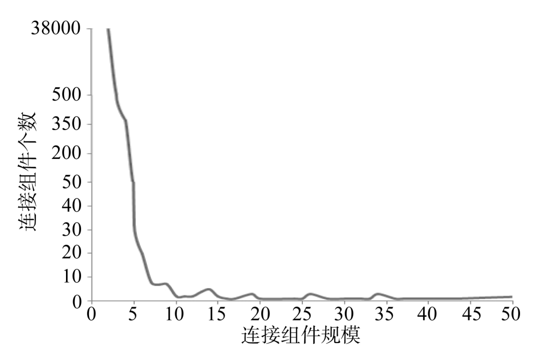


图 2　连接组件规模分布

(2) 节点的度数、入度和出度

度是描述节点属性的重要概念。在网络中，节点 v_i 的邻边数目 k_i 称为该节点的度。一个节点的度越大，该节点越重要。对网络中所有节点的度求平均，可得到网络的平均度。有向网络中与某个节点相连的线既有指向节点的，也有从节点发出的，因此也有必要分开统计两个方向的连线数，前者称为节点的入度，后者称为出度。在社交网络中，通常将入度视为声望，将出度视为合群性[10]。

样本 sameAs 网络中，节点度数分布如图 3 所示。98%的节点的度数为 1，即只与一个节点进行 sameAs 关联，分布尾部稀少，少量的节点与较多的 RDF 资源进行了 sameAs 关联。节点度数分布图在头部呈现出指数行为，尾部呈现长尾特征。样本 sameAs 网络的结构特征表明即使个别节点失效，不至于影响整体的稳定性，但高度数节点失效，会对关联数据网络造成一定的影响。

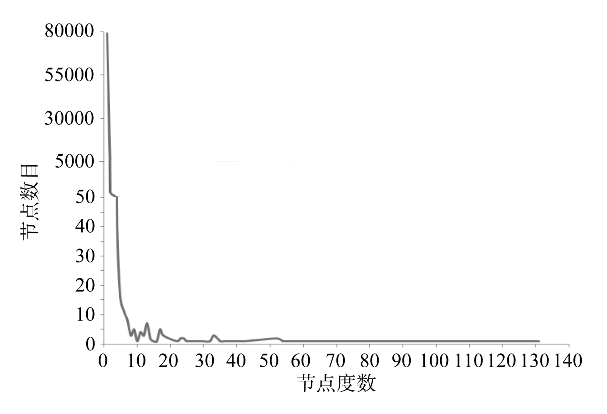


图 3　节点度数分布

节点的入度范围在 0-7 之间，48.6%的节点入度为 0，50.7%的节点入度为 1。节点的出度范围在 0-131 之间，51.3%的节点出度为 0，47.6%的节点出度为 1。两个 RDF 资源间大多数为单向连接，只有极少数 RDF 资源间建立了双向连接。RDF 资源更容易与其他数据集中的资源主动建立 owl:sameAs 连接。正如 Vatant 指出，当 owl:sameAs 用于数据融合时未必是对称属性。假设 A 拥有资源 a，B 拥有资源 b，“a owl:sameAs b”并不意味着“b owl:sameAs a”。只有 A 声明了“a owl:sameAs b”，B 也声明了“b owl:sameAs a”，a 和 b 这两个 RDF 资源才被认为有强等同关系[12]。在样本 sameAs 网络中，RDF 资源间建立了双向连接的情况非常少，基于 sameAs 进行语义聚合时要适当考虑此类情况。

3.2　sameAs 网络域名分析

顶级域名通常可以用来识别关联数据的发布者，即资源的拥有者(即拥有 URIs 命名空间并且对相关 URIs 作出官方描述的责任人)。通常假设顶级域名相同的数据属于一个数据集[3]。对于单个域名下包含多个数据集的情况，单独处理这部分数据。研究顶级域名之间的连接情况可以发现数据集之间的关联情况。从这些 sameAs 陈述中提取所有 RDF 资源的域名，进而统计分析不同数据集之间的 owl:sameAs 连接情况。经过对 80 521 个资源 URIs 的提取，得到 136 个不同顶级域名。利用 Gephi 绘制基于实例 sameAs 连接的域名网络结构图，并通过特定的布局工具对节点进行类聚和排列，最终效果如图 4 所示。在该图中，不同节点代表不同的数据集，节点颜色的深浅代表入度的大小，节点的大小代表出度的大小，有向连线代表数据集之间的 sameAs 连接，连线的粗细代表连接的权重。发出有向连线的节点一方称为源数据集，有向连线指向的节点一方称为目标数据集。

从图 4 中可以看到不同数据发布者之间的联系：SEC Edgar (edgarwrap.ontologycentral.com)同 Freebase (rdf.freebase.com)建立了密集的 sameAs 连接，DBTune (dbtune.org，提供音乐相关的结构化数据)和 BBC (bbc.co.uk)次之，DrugBank (wifo5-04.informatik.uni-mannheim.de/drugbank/，药物库)和 LinkedCT (data.linkedct.org，临床试验关联项目)之间也建立了数量可观的 sameAs 连接。笔者认为彼此之间建立了大量 sameAs 连接的域名，从不同角度描述了相似的话题。利用 Gephi 中的布局工具，把性质相同的节点聚在一起并从整体上作有序排列，有利于进一步发现享有共同知识和兴趣的数据发布者。在图 4 中可以看到一些比较大的簇，如以 DBpedia 为中心、以 BibSonomy 为中心的簇。DBpedia 与许多大规模的数据集和本体实现关联和互操作，而由于 DBpedia 广泛的主题覆盖，因此它也被各种数据集首选作关联目标。在样本 sameAs 网络中，进一步验证了 DBpedia 被称为“关联中转站”这一事实[13]。BibSonomy 是由 Kassel 大学中

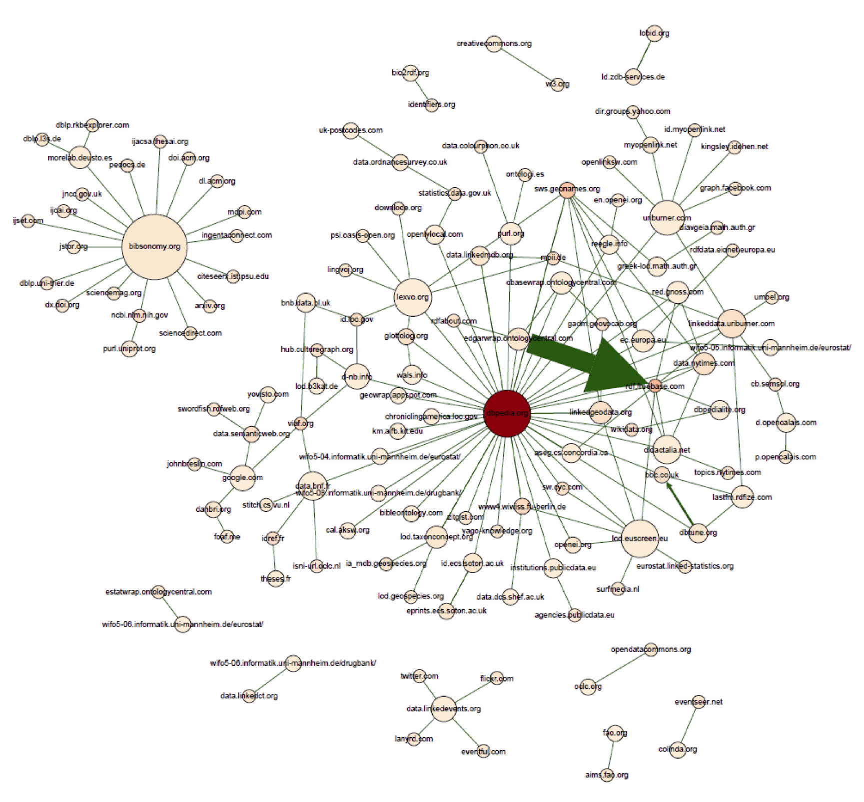


图 4　样本 sameAs 网络域名网络结构图

知识和数据工程组研究的用于共享标签和文献列表的推荐系统，旨在整合书签系统和团队出版物管理的特征,使用户能够储存和组织标签及发布的条目[14]。BibSonomy 通过提供文献交流的社会平台，支持不同社区和用户合作。以 BibSonomy 为中心的簇代表一个社区，其成员 ACM 数字图书馆(http://dl.acm.org/)等发布关于学术期刊及文献的信息，DBLP (http://dblp.uni-trier.de/)等提供关于计算机科学期刊和论文集的开放书目信息，NCBI(https://www.ncbi.nlm.nih.gov/)作为国家生物技术信息中心发布相关科学研究数据。由于本文数据是真实数据网络中的一部分，在样本数据中，BibSonomy 作为中心点，与之关联的数据集较多，但其与每个关联数据集之间的 sameAs 陈述并不多。也正是由于这个原因，虽然在 LOD 云图中，DBLP、NCBI 等数据集与 DBpedia 都是有连接的，但由于样本数据中恰好没有这部分 sameAs 陈述，因此在图 4 中没有看到上述数据源与 DBpedia 的连接。另外，还有一些比较小但有意思的簇，如以 EUscreen (http://lod.euscreen.eu/)为中心的簇。EUscreen 旨在创造欧洲电视节目、二次资源及文章的收集，以便学生、学者和普通大众获取使用[15]，因此其以关联数据的形式发布相关内容，使用户不仅能通过标准网络技术获取和检索相关元数据，而且能发现更多相关的可用数据，进而通过应用程序集成 EUscreen 收集的数据。

为深入了解数据集之间的连接情况，对每个数据集的入度和出度进行统计并比对。如图 5 所示，蓝色的线代表数据集的入度，红色的线代表数据集的出度。发现高度链接的数据集较少且其入度和出度相差较大，大部分数据集度数较低、owl:sameAs 连接稀疏，部分数据集只有入度或出度(即要么被动关联要么主动关联)。

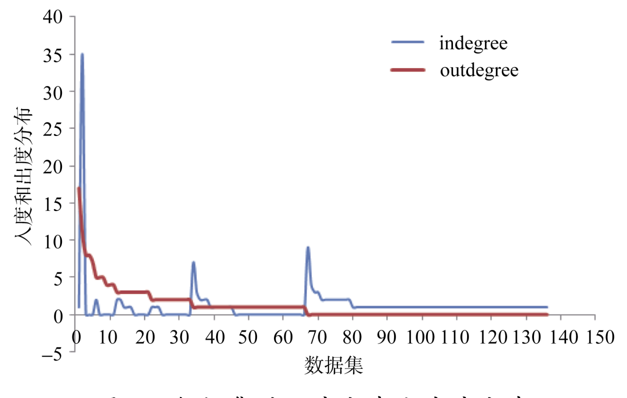


图 5　数据集的入度分布和出度分布

3.3 基于 owl:sameAs 的实例类型分析

了解真实数据网络中建立了 owl:sameAs 连接的实体对象类型有助于探寻关联数据的分布和应用领域，从而开拓不同领域的关联发现和应用。因此，基于 owl:sameAs 连接，对源数据集中实例的 rdf:type 信息进行提取并统计分析。由于数据发布者可以从不同维度描述同一实体对象，因此同一实体常对应多个类型。在源数据集中，总共获得 5 155 个 RDF 资源的 9 056 个类型信息。其中有 340 个实体对象对应的类型数目大于 1, DBpedia 中的足球运动员艾度斯恩(Connally Edozien)所属的类型更是多达 65 个，其所属类型从不同角度描述了同一个人。为避免多次重复统计同一个实体对象，对于类型数目大于 1 的实体只取其中一个类型(并不影响其最终归并后的类型)，经过分组汇总最终获得 181 个有效的以 HTTP 形式表示的不同类型信息及其对应的实体个数。基于类型查看建立了 owl:sameAs 连接的实体对象类型并再次归并及统计其数目，结果如图 6 所示(对于拥有实例数目小于 20 的类型在此处不作讨论)。可以看到关联数据网络中，建立了最多 owl:sameAs 连接的实体对象类型为人名，其次分别是地名、医药类名称、机构名称、电影等。

近几年，利用实例数据进行概念对齐显示出其有效性。Parundekar 等[5]提出识别包含于概念的等同实例将会导致这些概念之间的对齐。同时对源数据集和目标数据集中由 owl:sameAs 连接的实例的 rdf:type 信息进行统计，有利于发现不同数据集中可对齐的概念，同时可以帮助了解不同数据集之间建立 owl:sameAs 连接的深层次原因。图 7 是对样本数据集中的一个 owl:sameAs 连接同时获取其 RDF 资源的类型信息。

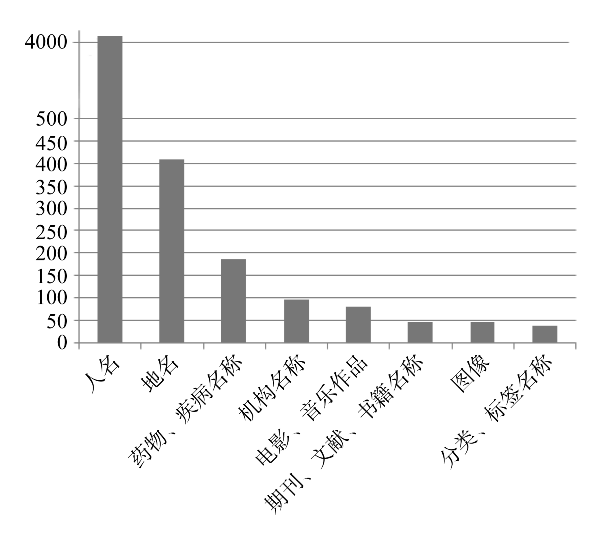


图 6 基于 owl:sameAs 连接的实体对象类型分布

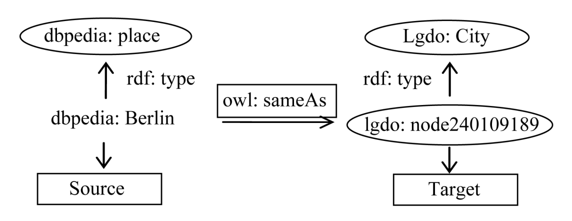


图 7 基于 owl:sameAs 连接的实例的 rdf:type 信息获取

从图 7 可以看到在 DBpedia 数据集中，柏林(dbpedia: Berlin)是 dbpedia:place 类的一个实例，在 linkedgeodata 数据集中 lgdo: node240109189 是 lgdo:City 类的一个实例，因而可知 dbpedia:place 和 lgdo:City 这两个概念至少是有交集的。在样本数据集中，基于 owl:sameAs 连接同时获取源数据集和目标数据集中实例的 rdf:type 信息，共计得到 625 条记录。基于这 625 条记录统计源数据集和目标数据集使用最多的 type 类型对，如表 3 所示。

表 3 基于 owl:sameAs 连接的源数据集和目标数据集常用 type

源数据集	目标数据集	基于 owl:sameAs 连接的最常用的 type 对	
		源数据集 type	目标数据集 type
theses.fr	idref.fr	<http://www.abes.fr/foafPerson>	<http://xmlns.com/foaf/0.1/Person>
		<http://www.abes.fr/foafAgent>	<http://xmlns.com/foaf/0.1/Person>
		<http://www.abes.fr/foafAgent>	<http://xmlns.com/foaf/0.1/Organization>
d-nb.info	dbpedia.org	<http://d-nb.info/standards/elementset/gnd#DifferentiatedPerson>	<http://dbpedia.org/class/yago/Traveler109629752>
morelab.deusto.es	dblp.l3s.de	<http://swrc.ontoware.org/ontology#Article>	<http://purl.org/dc/dcmitype/Text>
wals.info	glottolog.org	<http://purl.org/dc/terms/LinguisticSystem>	<http://purl.org/linguistics/gold/Language>
didactalia.net	data.nytimes.com	<http://rdfs.org/sioc/types#Tag>	<http://www.w3.org/2004/02/skos/core#Concept>

表 3 中，通过第 1、2、3、4 行可以看出：基于 owl:sameAs 实例连接，源数据集和目标数据集可以尝试进行对齐的概念有哪些，这为将来不同数据集的概念之间的对齐提供了有益的参考。第 5 行中对于 nytimes 数据集，虽然其包含丰富的术语层次，但只涵盖了很少的概念，大部分实体归属于 skos:Concept 这个概念，因此该数据集与其他数据集进行本体对齐时提供的概念非常有限。

4 讨 论

sameAs 网络特征表明大部分节点只有一个 owl:sameAs 连接，少数节点拥有多个甚至大量的 owl:sameAs 连接。现实网络具有优先连接的特征，即新的节点更倾向于与那些具有较高度的“大”节点相连接，sameAs 网络同样具有这种特征。对于 sameAs 网络而言，大部分节点随机失效基本不会影响其连通性，但少数重要节点的失效就会对网络的连通性造成一定影响，进而影响数据的关联。sameAs 网络连接组件规模较小(典型尺寸为 2)，不利于数据集之间关联关系的扩散。有时数据发布者并不热衷于声明 owl:sameAs 连接也会影响到连接组件的规模。

对基于 owl:sameAs 实例连接的部分数据集的入度和出度进行统计，发现综合类知识库(如 DBpedia, Freebase)及同类领域中的知名数据集(如地理领域中的 GeoNames)容易被其他数据集信任并作为链接资源，因此这些数据集往往具有高入度。其中一些数据集的入度和出度相差较大，如 DBpedia。入度高说明其作为知名数据集由于内容跨度大而被后发布的数据集积极关联，而出度小则说明其发布较早且后期维护滞后，致使其未能与后发布到 LOD 中的数据集主动关联，从而减少了数据集之间的互联。跨领域中的语言资源 Lexvo 数据集具有高出度而无入度，这是由于其需要确保所发布的资源即有关语言的实体对象可以与网络中多样化的资源建立密集的关联，因此它与较多数据集的实体对象主动建立了 owl:sameAs 关联。BibSonomy 数据集与之类似，出度远远超过入度，说明其作为分享标签和文学作品的推荐系统基于自身属性从而积极与有着类似话题的数据集建立关联，而被其关联的数据集大多是出版物领域较知名权威的期刊或科研组织，由于发布数据集的出发点不同、时间先后不同、发布者的地位不同等原因，在其之间未能建立对等连接。在整个数据网络中，对不同数据集中同一实体的关联发现还有很大的开拓空间。

5 结 语

互连数据集通常具有互补数据，某一实体的事实可能分布于若干数据集，将同一实体的不同属性及属性值聚合可以产生基于不同观点的实体的完整呈现。因此，owl:sameAs 连接在数据集互联中起着举足轻重的作用。sameAs 网络结构具有连接组件规模较小，高度数节点稀疏，大部分节点连接单一化，节点出、入度分布曲线具有在头部呈幂率分布、尾部呈长尾分布的特征。基于 owl:sameAs 连接的关联数据集大部分连接稀疏，高度链接的数据集较少且其中部分出入度相差较大。LOD 云图中的数据集大部分基于实例对齐技术，通过实例级关系相互连接，而基于实例的 owl:sameAs 连接可以进行概念对齐、属性对齐等，从而实现本体对齐。本体对齐通过为数据聚合、跨数据集查询、知识获取提供解决方案，从而使 LOD 数据集的事实和信息呈现更加有用。在数据网络中，从不同数据集中找到“等同”实例是有挑战性的，发现分布于不同数据集的等同实体并为之建立 owl:sameAs 连接，需要进一步提高相关技术、完善关联机制。owl:sameAs 属性是否具有对称性、传递性、适用条件以及 owl:sameAs 属性在推理中的应用机制，这些问题需要在今后进一步研究，它们的应用势必会改变 sameAs 网络的结构。

参考文献：

[1] Bizer C, Tom H, Berners-Lee T, et al. Linked Data: The Story So Far [J]. International Journal on Semantic Web & Information Systems, 2009, 5(3): 1-22.

[2] Abele A, McCrae J. Linking Open Data Cloud Diagram 2017 [EB/OL]. [2017-03-07]. http://lod-cloud.net/.

[3] Schmachtenberg M, Bizer C, Paulheim H. Adoption of the Linked Data Best Practices in Different Topical Domains [C]// Proceedings of the 13th International Semantic Web Conference. 2014: 245-260.

[4] Gunaratna K, Lalithsena S, Sheth A. Alignment and Dataset Identification of Linked Data in Semantic Web[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014, 4(2): 139-151.

[5] Parundekar R, Knoblock C A, Ambite J L. Linking and Building Ontologies of Linked Data[C]// Proceedings of the 9th International Semantic Web Conference, Shanghai, China. 2010.
[6] Correndo G, Penta A, Gibbins N, et al. Statistical Analysis of the owl:sameAs Network for Aligning Concepts in the Linking Open Data Cloud[J]. Lecture Notes in Computer Science, 2012, 7447(5): 215-230.
[7] Nikolov A, Motta E. Capturing Emerging Relations Between Schema Ontologies on the Web of Data[C]//Proceedings of the 9th Semantic Web Conference, Shanghai, China. 2010.
[8] Gunaratna K, Thirunarayan K, Jain P, et al. A Statistical and Schema Independent Approach to Identify Equivalent Properties on Linked Data[C]// Proceedings of the 9th International Conference on Semantic Systems. ACM, 2013: 33-40.
[9] Bechhofer S, van Harmelen F, Hendler J, et al. OWL Web Ontology Language Reference [EB/OL]. [2016-11-02]. https://www.w3.org/TR/owl-ref/#sameAs-def.
[10] 郭世泽, 陆哲明. 复杂网络基础理论[M]. 北京: 科学出版社, 2012. (Guo Shize, Lu Zheming. Basic Theory of Complex Networks [M].Beijing: Science Press, 2012.)
[11] Tobias K, Andreas H. Billion Triples Challenge 2014 Dataset [EB/OL]. [2016-10-11]. http://km.aifb.kit.edu/projects/btc-2014/.
[12] Using owl:sameAs in Linked Data[EB/OL]. [2016-10-12]. http://blog.hubjects.com/2007/07/using-owlsameas-in-linked-data.html.
[13] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data[C]// Proceedings of the 6th International Semantic Web Conference on Semantic Web. 2007.
[14] Hotho A. BibSonomy: A Social Bookmark and Publication Sharing System[C]// Proceedings of the 14th International Conference on Conceptual Structures, Aalborg, Denmark. Aalborg University Press, 2006.
[15] EUscreen Linked Open Data Pilot [EB/OL]. [2017-03-08]. http://lod.euscreen.eu/.

作者贡献声明:

贾君枝: 提出研究思路, 设计研究方案, 修改论文;
李晓: 收集、整理、分析资料, 撰写论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 李晓. sample.txt. 样本数据集.

收稿日期: 2017-05-05
收修改稿日期: 2017-07-12

Analyzing owl:sameAs Network in Linked Data

Jia Junzhi Li Xiao
(School of Economics and Management, Shanxi University, Taiyuan 030006, China)

Abstract: [**Objective**] This paper examines the application of the owl:sameAs link in the Web of Data. [**Methods**] First, we extracted owl:sameAs links from the BTC 2014 dataset. Then, we analyzed the structure of the sample data, as well as their domain names and instance types. [**Results**] The retrieved links of owl:sameAs were sparse, and most entities only had single connection between each other. [**Limitations**] The size of our sample data was small, and more comprehensive analysis was needed. [**Conclusions**] Our study lays some foundations for data integration, ontology alignment, knowledge discovery of the Web of Data.

Keywords: owl:sameAs Interlinking of Datasets Network